(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開平7－13959
(43)公開日　平成７年(1995)１月17日

(51)Int.Cl.[6]	識別記号	庁内整理番号	FI	技術表示箇所
G06F 17/00				
G01N 33/68		7055－2J		
G06F 17/30				
		8724－5L	G06F 15/20	D
		9194－5L	15/40	530 S

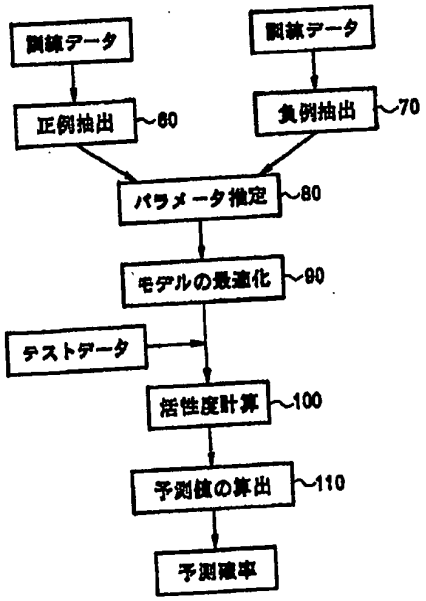
審査請求　有　請求項の数６　ＯＬ　（全９頁）

(21)出願番号　特願平4－124817
(22)出願日　平成４年(1992)５月18日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　馬見塚　拓
東京都港区芝五丁目７番１号日本電気株式会社内
(72)発明者　山西　健司
東京都港区芝五丁目７番１号日本電気株式会社内
(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】　タンパク質立体構造予測方法

(57)【要約】
【目的】　構造未知のタンパク質のアミノ酸配列情報から、それに対応する二次構造（αヘリックス）を高い信頼性で予測する。
【構成】　ステップ６０で構造既知及び未知のタンパク質アミノ酸配列を入力とし、それらのアライメント（整合）から二次構造領域の正例を出力し、ステップ７０で構造既知のタンパク質アミノ酸配列を入力とし、それらのアライメントから二次構造領域の負例を出力し、ステップ８０で前記正例と負例を入力とし、これら訓練データのアミノ酸の実数値属性から確率的規則の実数値パラメータの推定値を出力し、ステップ９０で情報量規準を用いて前記確率的規則の最適化したセル数を出力とし、ステップ１００でテストデータ配列を入力とし、前記確率的規則を使用し、計算を行ったテストデータ配列の各領域に対する活性度を出力し、ステップ１１０で前記活性度を入力とし、その中から最適値を出力する。



1

【特許請求の範囲】

【請求項1】　タンパク質のアミノ酸配列からタンパク質の構造予測を行うための訓練データを抽出する訓練データ抽出手段と、訓練データから確率的規則の学習を行う学習手段と、学習された確率的規則を用いてテストアミノ酸配列データに対してテストを行うテスト手段とから成ることを特徴とするタンパク質立体構造予測方法。

【請求項2】　前記訓練データ抽出手段が、構造既知のタンパク質のアミノ酸配列に対して、同じファミリーに属するタンパク質のアライメント（整合）をとり、予測対象とする二次構造領域に対応する部分配列を、二次構造領域の正例として抽出するステップと、構造既知のタンパク質の予測対象とする二次構造に対応する部分配列に対して、構造既知のタンパク質からなるデータベースの各配列のアライメントをとり、予測対象とする二次構造に対応しない部分配列を、二次構造領域の負例として抽出するステップとから成ることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項3】　前記学習手段が、前記正例と前記負例とからなる学習データのアミノ酸の種類から、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定することを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項4】　前記学習手段が、前記正例と前記負例とからなる学習データのアミノ酸の実数値属性から、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定することを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項5】　前記学習手段が、前記正例と前記負例とからなる学習データのアミノ酸の実数値属性から、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定するステップと、前記ステップの確率的規則におけるモデルを情報量規準を用いて最適化するステップとから成ることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項6】　前記テストデータ配列に対するテスト手段が、前記学習方法により学習された確率的規則を使用し、テストデータ配列の各領域に対して、その活動度を計算するステップと、計算された活動度の中から最適値を選出するステップとから成ることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、構造未知のタンパク質アミノ酸配列から、そのタンパク質の立体構造を予測する方法に関する。

【0002】

【従来の技術】タンパク質のアミノ酸配列情報を用いて、そのタンパク質内の立体構造を予測する問題の一つとして、タンパク質二次構造予測問題がある。二次構造

2

とは、αヘリックスやβシートといったタンパク質内部でのまとまりのある構造を指し、二次構造予測問題は、タンパク質のアミノ酸配列情報を用いて、3（あるいは4）種類の二次構造の中から、一次配列の各残基（以下、予測対象となる残基を中心残基とする）に対応する一つの二次構造を予測する問題であり、二次構造予測が可能になることにより、タンパク質の立体的な構造予測も可能になると考えられている。図3は、本発明の二次構造（αヘリックス）領域予測方法を示す模式図であるが、従来技術によるタンパク質の二次構造を予測する方法として、例えば、1974年発行の米国の雑誌「バイオケミストリー」（Biochemistry）の第23巻222－245頁掲載のチョウ（Chou）とファスマン（Fasman）による論文「プレディクション オブ プロテイン コンホメーション」（Prediction of protein conformation）（以下、CF法と略す）、1978年発行の米国の雑誌「ジャーナルオブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第120巻97－120頁掲載のガルニエ（Garnier）らによる論文「アナリシス オブ ザ アキュレシー アンド インプリケーションズ オブ シンプル メソード フォー プレディクティング ザ セコンダリー ストラクチャー オブ グロブラー プロテインズ」（Analysis of the accuracy and implications of simple method for predicting the secondary structure of globular proteins）（以下、GOR法と略す）、1987年発行の米国の雑誌「ジャーナル オブ モレキュラ バイオロジー」（Jouranal of Molecular Biology）の第198巻425－443頁掲載のギブラト（Gibrat）らによる論文「ファザー デベロプメンツ オブ プロテイン セコンダリー ストラクチャ プレディクション ユージング インホメーション セオリー：ニュー パラメータズ アンド コンシダレーション オブ レジデュー ペアズ」（Further developments of protein secondary structure prediction using information theory:New parameters and consideration ofresidue pairs）（以下、GGR法と略す）、及び1988年発行の米国の雑誌「ジャーナル オブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第202巻865－884頁掲載のキャン（Qian）らによる論文「プレディクティング ザ

セコンダリー　ストラクチャー　オブ　グロブラー　プロテインズ　ユージング　ニューラル　ネットワーク　モデルズ」（Predicting the secondary structure of globular proteins using neural network models）（以下、ＱＳ法と略す）などがある。ＣＦ法は、タンパク質構造のデータベースから各二次構造におけるアミノ酸の統計的な出現頻度を求め、この頻度表を使用し、経験的な規則に基づく予測を行っている。また、ＧＯＲ法は、中心残基の二次構造に対して、その残基から数残基離れた残基により独立にもたらされる情報量の和を計算し、その相対値から予測を行い、ＧＧＲ法は、中心残基の二次構造に対して、その残基及びその残基から数残基離れた残基によりもたらされる情報量の和から予測を行っている。さらに、ＱＳ法は、３層のフィードフォワード型のネットワークを使用し、中心残基の前後８残基を含む配列を入力とし、二次構造に対する中心残基及び周辺残基からの寄与をニューラルネットワークを用いて抽出することにより予測を行っている。

【０００３】

【発明が解決しようとする課題】３種類の二次構造の中からアミノ酸配列の各残基に対応する二次構造を選択する予測を３状態予測と呼ぶが、その予測結果の尺度である予測率は、従来の技術のいずれの方法も３状態予測で６０％台であり、αヘリックスにだけ限ってより予測率の高い予測手法が望まれていた。また、従来の予測結果は、アミノ酸一次配列内の各中心残基に対応する二次構造を予測する残基対応の予測であり、一次配列内のどの領域がどの二次構造に相当するかといった領域対応の予測を行うことも重要であるにも関わらず、このような予測方式に十分な検討がなされていなかった。さらに、アミノ酸配列を文字列としてのみならず、そのアミノ酸の性質（疎水性、分子量など）を考慮した予測を行うことによる予測法も全く確立されていなかった。

【０００４】

【課題を解決するための手段】第１の発明は、タンパク質のアミノ酸配列からタンパク質の構造予測を行うための訓練データを抽出する訓練データ抽出手段と、訓練データから確率的規則の学習を行う学習手段と、学習された確率的規則を用いてテストアミノ酸配列データに対してテストを行うテスト手段とから成ることを特徴とする。

【０００５】第２の発明は、前記訓練データ抽出手段が、構造既知のタンパク質のアミノ酸配列に対して、同じファミリーに属するタンパク質のアライメント（整合）をとり、予測対象とする二次構造領域に対応する部分配列を、二次構造領域の正例として抽出するステップと、構造既知のタンパク質の予測対象とする二次構造に対応する部分配列に対して、構造既知のタンパク質からなるデータベースの各配列のアライメントをとり、予測対象とする二次構造に対応しない部分配列を、二次構造領域の負例として抽出するステップとから成ることを特徴とする。

【０００６】第３の発明は、前記学習手段が、前記正例と前記負例とからなる学習データのアミノ酸の種類から、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定することを特徴とする。

【０００７】第４の発明は、前記学習手段が、前記正例と前記負例とからなる学習データのアミノ酸の実数値属性から、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定することを特徴とする。

【０００８】第５の発明は、前記学習手段が、前記正例と前記負例とからなる学習データのアミノ酸の実数値属性から、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定するステップと、前記ステップの確率的規則におけるモデルを情報量規準を用いて最適化するステップとから成ることを特徴とする。

【０００９】第６の発明は、前記テストデータ配列に対するテスト手段が、前記学習方法により学習された確率的規則を使用し、テストデータ配列の各領域に対して、その活動度を計算するステップと、計算された活動度の中から最適値を選出するステップとから成ることを特徴とする。

【００１０】

【実施例】次に、本発明について図面を参照して詳細に説明する。

【００１１】図１は、本発明のタンパク質立体構造予測方法の実施例を説明するフローチャートである。本実施例では、対象とする二次構造としてαヘリックスを扱うものとする。

【００１２】ステップ１０は、第２の発明に含まれる。このステップでは、αヘリックスの領域がわかっているタンパク質のアミノ酸配列に対して、同じファミリーのタンパク質、例えば、種が異なる同じタンパク質のアライメント（整合）をとり、αヘリックスに対応する部分配列を、αヘリックスの正例として抽出する。

【００１３】例えば、ヘモグロビンというタンパク質のβ鎖の場合には、ヒトのヘモグロビンのαヘリックスの位置は、Ｘ線結晶回折の結果から明らかになっており、８個のαヘリックスの領域を有することが知られている。従って、ヒトのヘモグロビンβ鎖に対して、他の種、例えば、チンパンジー、ウマなどの他の種のヘモグロビンβ鎖のアライメントをとり、８個のαヘリックスに対応する領域をαヘリックスの正例として抽出する。

【００１４】ステップ２０は、第２の発明に含まれる。このステップでは、αヘリックス位置の知られているタンパク質のαヘリックスに対応する部分配列に対して、αヘリックス位置の知られているアミノ酸配列データベースの各配列のアライメントをとり、αヘリックスに対

応しない部分配列を、ステップ１０で抽出されたαヘリックスの正例に対する負例として抽出する。

【００１５】ヘモグロビンβ鎖の例では、８個のαヘリックスに対応する部分配列に対して、例えば、ＰＤＢ（Ｐｒｏｔｅｉｎ Ｄａｔａ Ｂａｎｋ）などのタンパク質構造データベース内のいくつかの配列に対してアライメントを行い、アライメントの結果得られた各部分配列において、その配列の構造がαヘリックスではない場合に、それらを負例として抽出する。例えば、負例抽出の際のアライメントでは、一定の割合以上の相同性を保持する部分配列を負例とすることが考えられる。具体的には、アライメントによる相同性が３０％以上の部分配列を負例とする方法などがある。

【００１６】抽出するデータ数については、例えば、αヘリックスの正例となる各領域における正例と負例との割合を各領域についてそれぞれ等しくすることが考えられ、また例えば、その割合として正例、負例を同数とすることが考えられる。

【００１７】ステップ３０は、第３の発明、第４の発明、第５の発明に共通に含まれ、確率的規則の実数値パラメータを推定するステップである。このステップでは、ステップ１０で求めた正例とステップ２０で求めた負例からなる学習データから、確率的規則を用いることにより、この確率的規則の実数値パラメータを推定する。このステップでの確率的規則の構造を、以下に示す。

【００１８】確率的規則とは、ここでは任意の与えられた配列の領域に対して、αヘリックスが対応する確率を与える確率分布のことである。各χ_i（ｉ＝１，…，ｎ）をそれぞれ属性値の空間として、χをそれらの直積、すなわち、$\chi=\chi_1\times\chi_2\times\cdots\times\chi_n$と書く。

【００１９】例えば、χは２０種類のアミノ酸からなる一つの集合を表す場合や、また$\chi=\chi_1\times\chi_2$で、χ_1が疎水性を表す数値の範囲かつχ_2が分子量を表す数値の範囲を表す場合などがある。この例での前者の場合が第３の発明で使用され、それ以外の場合が第４の発明及び第５の発明で使用される。Ｓをある領域の長さＷの配列であり、各Ｓは$\chi\times\chi\times\cdots\times\chi$の元とみなし、また、$X_i$を配列Ｓの左から数えてｉ番目の残基であり、$P(\alpha\mid X_i)$が、$X_i$に対応する二次構造がαヘリックスである確率とする。ここで、配列Ｓに対応する二次構造がαヘリックスである確率$P(\alpha\mid S)$は、$P(\alpha\mid X_i)$の積として次のようにかけるものと仮定する。

【００２０】

$$P(\alpha\mid S)=\Pi_{i=1}^{w}P(\alpha\mid X_i)$$

さらに、各$P(\alpha\mid X_i)$の具体的表現として、例えば、有限分割型確率的規則を使用する。有限分割型確率的規則は次のような構造をもつ条件付き確率分布であり、以下のように構成する。前記配列Ｓのｉ番目の残基における属性の実数値のとり得る範囲を重なり合わない部分領域（以下、これをセルと呼ぶ）に分割し、ｍを全セル数、C_kをｋ番目のセルとした時に、X_iがｍ個のセルの内のC_kに含まれる場合に、$P(\alpha\mid X_i)=P_k(i)$とする。ここで、

【００２１】

【数１】

$$p_k(i)\in[0,1]\ (k=1,\cdots,m)$$

【００２２】であり、これを確率パラメータと呼ぶ。図４は、有限分割型確率規則の構造を示す模式図であるが、この図では、一例として、値が０から１の範囲をとる一つの属性により確率パラメータを推定する場合を示す。

【００２３】確率パラメータは、各セルに含まれる正例及び負例のデータ数を用いて推定する。ｍをセルの数、$N_k^+(i)$をｉ番目の位置でのｋ番目のセルに含まれる正例数、$N_k^-(i)$をｉ番目の位置でのｋ番目のセルに含まれる負例数、$N_k(i)$をｉ番目の位置でのｋ番目のセルに含まれる正例数と負例数の和とし、ｉ番目の位置でのｋ番目のセルにおける推定値を

【００２４】

【数２】

$$\hat{p}_k(i)$$

【００２５】とする。例えば、次式のラプラス推定量によって、各セルに対する確率パラメータを計算する。

【００２６】

【数３】

$$\hat{p}_k(i)=(N_k^+(i)+1)/(N_k(i)+2)$$

$$(k=1,\cdots,m)$$

【００２７】ただし、推定量はラプラス推定量のみならず、多くの推定量が使用できる。

【００２８】ステップ４０は、第６の発明に含まれる。このステップでは、ステップ３０において学習された確率的規則を使用し、テストデータ配列の各領域に対して、その活性度を計算する。

【００２９】ここでは、活性度として尤度を使用する。

【００３０】具体的には、確率的規則が構成された長さｗのあるαヘリックス領域を考える。テストデータのアミノ酸配列に対して、前記領域の長さｗより小さな長さｔのｗ－ｔ＋１個のすべての部分領域を設け、このｗ－ｔ＋１個の部分領域それぞれをテストアミノ酸配列の左から順にあてはめていき、テスト配列の各領域の尤度を計算する。

【００３１】さて、ｋ番目の長さｔの部分領域に対して、αヘリックス領域の確率パラメータを左から順に並

べたものを$\xi_k = (\theta_1, \cdots, \theta_t)$，$\theta_i = (P_1(i), \cdots, P_m(i))$ $(i=1, \cdots, t)$と書く。

【0032】ここで、mはセルの数であり、θ_iは既に学習によって値が求められている。

【0033】w－t＋1個の部分領域の位置に対応して、このmt次元パラメータは、w－t＋1個求められるので、それを$\xi_1, \cdots, \xi_{w-t+1}$とする。

【0034】前記パラメータを使用して、任意の長さtのテストアミノ酸配列Γに対して、尤度がw－t＋1通り次のように計算できる。

【0035】

$P(\alpha \mid \Gamma : \xi_k)$ $(k=1, \cdots, w-t+1)$

ただし、各kについて、

$P(\alpha \mid \Gamma : \xi_k) = \Pi_{i=1}^{t} P(\alpha \mid \Gamma : \theta_i)$

ここで、$P(\alpha \mid \Gamma : \theta_i)$は$X_i$が1番目のセルに入れば、$P_1(i)$ $(l=1, \cdots, m)$と計算する。また、$P_I(i)$ $(l=1, \cdots, m)$はすでに学習されている。

【0036】例えば、前記有限分割型確率的規則でのアミノ酸の属性値の空間がある一つの属性値のみからなる場合で、またセルの数が3であり、セルに入る推定量はラプラス推定量により求めるとする。このとき、あるαヘリックス領域のk番目の長さ5の部分領域のi番目の位置での1番目のセルの正例数を$N_1^+(i)$、1番目のセルの負例数を$N_1^-(i)$、正例数と負例数の和を$N_1(i)$とする。すると、i番目の位置での1番目のセルの推定量は、例えば、$P_1(i) = (N_1^+(i)+1) / (N_1(i)+2)$として得られる。

【0037】ここで、テストアミノ酸配列のウィンドウの大きさ5の領域Γに対してテストを行うとし、領域Γのそれぞれの残基は前記部分領域での構成された学習規則での各1，2，3，2，1番目のセルに入る属性の実数値を有するとする。すると、前記k番目の部分領域によるテストアミノ酸配列の領域Γの活性度は尤度$P(\alpha \mid \Gamma : k)$として、次式のように計算される。

$$P(\alpha \mid \Gamma : k) = \{(N_1^+(1)+1)/(N_1(1)+2)\} \{(N_2^+(2)+1)/(N_2(2)+2)\} \{(N_3^+(3)+1)/(N_3(3)+2)\} \{(N_2^+(4)+1)/(N_2(4)+2)\} \{(N_1^+(5)+1)/(N_1(5)+2)\}$$

w－t＋1個の部分領域により、テストアミノ酸配列のとり得るすべての領域に対して、この尤度計算を行う。また、αヘリックス領域が複数個存在すれば、その各領域について同様の尤度計算を行う。

【0038】従って、テストアミノ酸配列内でのウィンドウの大きさに対応したすべての領域に対して、尤度が出力として得られることになる。

【0039】以上のウィンドウを使用した各領域に対応する尤度計算により、一つ一つの残基に対してαヘリックスが対応する確率を計算するのではなく、テストアミノ酸配列の各部分領域にαヘリックスが対応する確率を尤度として計算することができる。

【0040】ステップ50は、第6の発明に含まれる。このステップでは、前記ステップ40により計算された複数の活性度の中で、Γに対して最適な一つの活性度を求め、さらにテストアミノ酸配列全体における活性度の変化を出力する。

【0041】ステップ40に引き続きここでは、活性度として尤度を使用する。

【0042】例えば、最適値$P(\alpha \mid \Gamma : \xi_k^*)$を以下で定める。

$P(\alpha \mid \Gamma : \xi_k^*) = max\{P(\alpha \mid \Gamma : \xi_1), \cdots, P(\alpha \mid \Gamma : \xi_{w-t+1})\}$.

αヘリックス領域が複数個あれば、各領域について、同じΓに対して同様な尤度計算を行ない、αヘリックス領域全体を通じて最大の尤度を最適値として選ぶ方法も考えられる。

【0043】さらに、テスト配列内の尤度が与えられた各領域において、最大の尤度を領域内の各残基の最適値とする、あるいは、領域内の各残基に対しては、その残基を含む領域の得られた尤度の平均を各残基の最適値とする、などの方法を使用し、テストアミノ酸配列全体に対する尤度の変化を出力する。

【0044】以上の図1における学習及び予測方法は、αヘリックス以外の二次構造予測についても適用できる。

【0045】図2は、本発明のタンパク質立体構造予測方法の実施例を説明するフローチャートである。本実施例では、対象とする二次構造としてαヘリックスを扱うものとする。

【0046】ステップ60は、図1のステップ10と同じ処理を行ないαヘリックス領域予測のために必要な正例を抽出する。

【0047】ステップ70は、図1のステップ20と同じ処理を行ないαヘリックス領域予測のために必要な負例を抽出する。

【0048】ステップ80は、図1のステップ30と同じ処理を行ない確率的規則を適用し、この確率的規則の実数値パラメータを推定する。

【0049】ステップ90は、第5の発明にのみ含まれる。このステップでは、確率的規則のモデルを情報量規準を用いて最適化する。使用する情報量規準としては、例えば、MDL（minimum description length）規準などが考えられる。

【0050】前記有限分割型確率的規則にあてはめると、MDL原理は、データ記述長と有限分割型確率的規則による記述長との和が最小である時に最適な確率的規則が構成されているとする。なお、MDL原理について

9

は、１９７８年発行の米国の雑誌「オートマティカ」（Ａｕｔｏｍａｔｉｃａ）の第１４巻４６５－４７１頁記載のリサネン（Ｒｉｓｓａｎｅｎ）による論文「モデリング バイ ショーテスト データ ディスクリプション」（Ｍｏｄｅｌｉｎｇ ｂｙ ｓｈｏｒｔｅｓｔ ｄａｔａ ｄｅｓｃｒｉｐｔｉｏｎ）に詳しく記載されている。

10

＊【００５１】前記有限分割型確率的規則に対するデータの記述長は、規則の対数尤度の負をとることによって、次のように計算できる。ただし、以下では対数の底はすべて２とする。

【００５２】

【数４】

＊

$$-\Sigma_i \Sigma_k \{N_k^+(i) \log p_k(i) + (N_k(i) - N_k^+(i)) \log (1-p_k(i))\}$$

【００５３】また、前記有限分割型確率的規則の記述長は、各確率パラメータp_k（ｉ）の推定値はおよそ$\log N_k$（ｉ）ビットで記述できるから、次のように計算できる。

【００５４】

【数５】

※

$$\Sigma_i \Sigma_k (\log N_k(i))/2.$$

【００５５】したがって、ＭＤＬ原理によれば、次式が最小になるセル数ｍの大きさを確率規則を構成する最適なセル数とする。

【００５６】

※【数６】

$$-\Sigma_i \Sigma_k \{N_k^+(i) \log p_k(i) + (N_k(i) - N_k^+(i)) \log (1-p_k(i))\} + \Sigma_i \Sigma_k (\log N_k(i))/2.$$

【００５７】ステップ１００は、図１のステップ４０と同じ処理を行ない、ステップ９０を使用してモデルが最適化された確率的規則を使用し、テストアミノ酸配列データの各領域に対して、その活性度を計算する。

【００５８】ステップ１１０は、図１のステップ５０と同じ処理を行ないステップ４０により求められた複数の活性度から、配列全体に対する活性度の変化を出力する。

【００５９】以上の図２における学習及び予測方法は、αヘリックス以外の二次構造予測についても適用できる。

【００６０】

【発明の効果】二次構造既知のタンパク質のアミノ酸配列情報から、二次構造未知のタンパク質の二次構造を従来技術に対して高い精度で予測することができる。特に、アルブミンのαヘリックス領域を７０％以上の高い精度で予測可能である。また、ＭＤＬ原理などの情報量規準によりモデルを最適化することにより、確率的規則の構造を理論的に最適化することが可能になる。

【図面の簡単な説明】

【図１】本発明のタンパク質立体構造予測方法の一実施例を示すフローチャート

【図２】本発明のタンパク質立体構造予測方法の一実施例を示すフローチャート

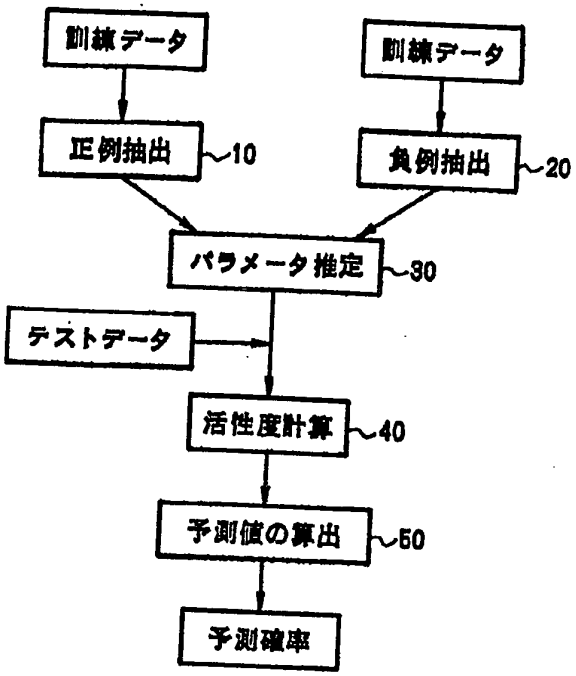
【図３】本発明の二次構造（αヘリックス）領域予測方法の模式図

【図４】本発明で使用する確率規則の一例である有限分割型確率規則の具体例を示す模式図

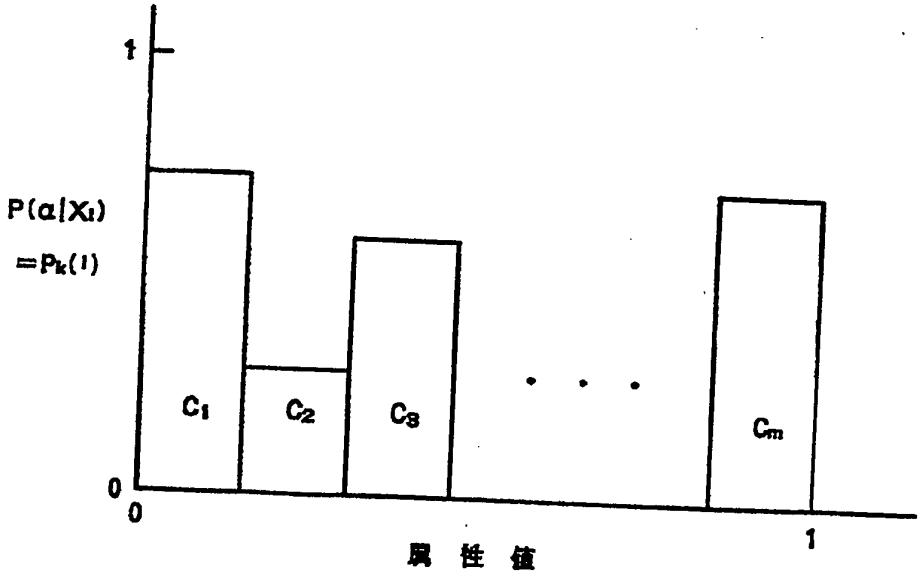
【符号の説明】

１０　正例抽出
２０　負例抽出
３０　確率的規則による実数値パラメータ推定
４０　テスト配列各領域に対する活性度計算
５０　テスト配列に対する予測値算出
６０　正例抽出
７０　負例抽出
８０　確率的規則による実数値パラメータ推定
９０　情報量規準による最適化
１００　テスト配列各領域に対する活性度計算
１１０　テスト配列に対する予測値算出

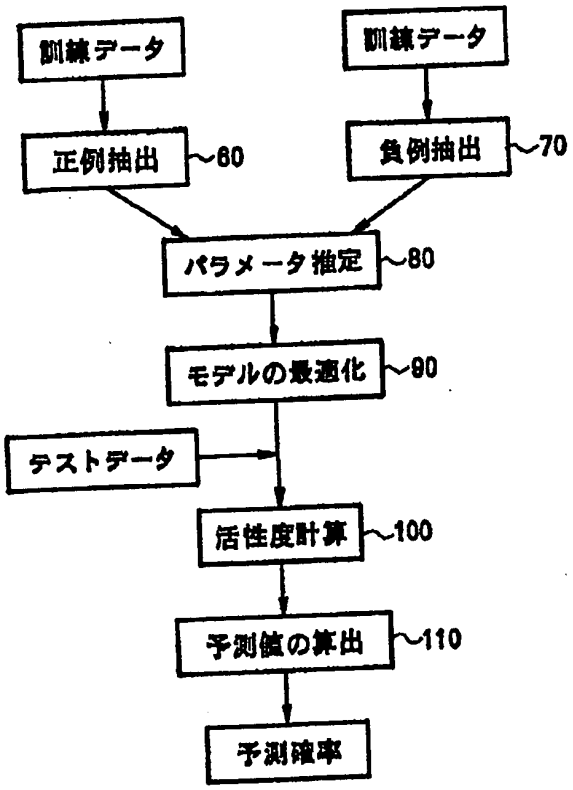
【図1】


訓練データ
訓練データ
正例抽出
10
負例抽出
20
パラメータ推定
30
テストデータ
活性度計算
40
予測値の算出
50
予測確率

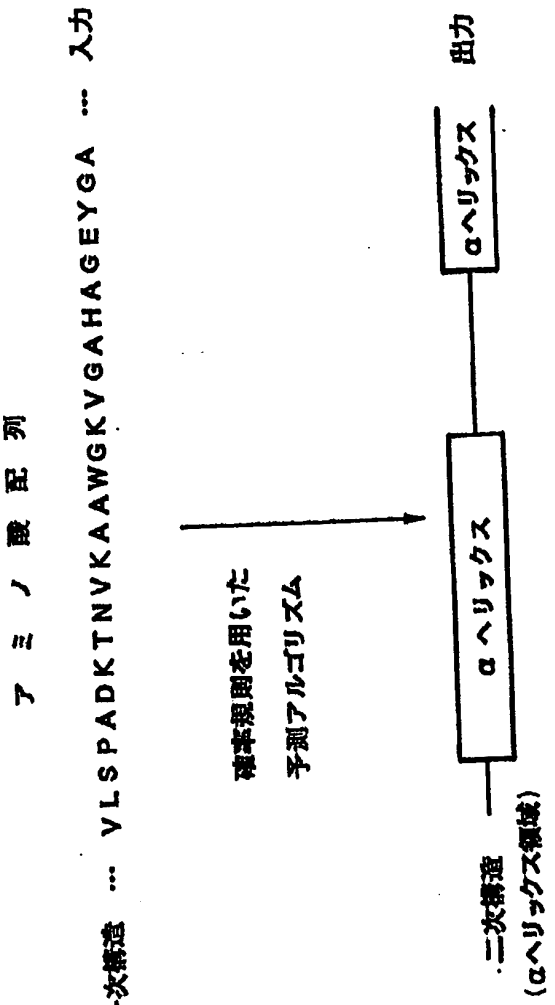
【図4】


1
P(α|Xi)
=Pk(1)
C1
C2
C3
Cm
0
0
1
属 性 値

【図2】


訓練データ
訓練データ
正例抽出
60
負例抽出
70
パラメータ推定
80
モデルの最適化
90
テストデータ
活性度計算
100
予測値の算出
110
予測確率

【図3】


アミノ酸配列
一次構造 … VLSPADKTNVKAAWGKVGAHAGEYGA … 入力
確率規則を用いた
予測アルゴリズム
αヘリックス
出力
二次構造
(αヘリックス領域)
αヘリックス